Definition

Project Overview

Finding a new customer is the main objective for the corporations in across various domains e.g. insurance, retail etc. With machine learning and high computing power we can process huge data and establish relations between the data and predict who will be the new customer

My passion with Python and finance made me to choose this customer segmentation project.

In this project, I analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population.

Use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, apply supervised learning technique on a dataset with demographics information for targets of a marketing campaign for the company and use a model to predict which individuals are most likely to convert into becoming customers for the company.

The data has been provided by Bertelsmann Arvato Analytics, and represents a real-life data science task. The data includes general population dataset, customer segment data set, mailout campaign dataset with response and test dataset that needs to make predictions.

Problem Statement

The goal is to find who will be the customer; the tasks involved are the following:

1. Data Exploration and Preprocessing
   Get to know the data and Identify missing or unknown data and impute the data accordingly.

2. Customer Segmentation using Unsupervised Learning
   Analyze general population and customer segment data sets and use supervised learning techniques to perform customer segmentation, identify the parts of the population the best describes the core customer base of the company.

   I will use Principal Component Analysis (PCA) technique for dimensionality reduction. Then, elbow curve will be used to identify the best number of clusters for KMeans algorithm. Finally, I will apply KMeans to make segmentation of population and customers and determine description of target cluster for the company.

3. Supervised Learning Model
   Build machine learning model using response of marketing campaign and use model to predict which individuals are most likely to convert into becoming customers for the company.

4. Kaggle Competition
   Submit results for Kaggle competition

Metrics

For imbalanced classification tasks like this, I am going to use Area under the receiver operating characteristic curve (ROC AUC) from predicted probabilities to evaluate performances of the models.

Analysis

Data Exploration

There are two data description excel sheets and four data files:

AZDIAS: Demographics data for the general population of Germany. It has 891211 persons data and 366 features.

```
In [4]: azdias.head(5)
```

Out[4]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTI
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0	11.(
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0	10.(
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0	1.(
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	3.(

5 rows × 366 columns

```
In [5]: azdias.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891221 entries, 0 to 891220
Columns: 366 entries, LNR to ALTERSKATEGORIE_GROB
dtypes: float64(267), int64(93), object(6)
memory usage: 2.4+ GB
```

CUSTOMERS: Demographics data for customers of a mail order company. It has 191652 rows and 369 features

```
In [20]: # View the first few lines of the customers dataframe
         customers.head()
```

Out[20]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	9626	2	1.0	10.0	NaN	NaN	NaN	NaN	10.0	1.(
1	9628	-1	9.0	11.0	NaN	NaN	NaN	NaN	NaN	NaN
2	143872	-1	1.0	6.0	NaN	NaN	NaN	NaN	0.0	1.(
3	143873	1	1.0	8.0	NaN	NaN	NaN	NaN	8.0	0.(
4	143874	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	7.(

5 rows × 369 columns

MAILOUT_TRAIN: Demographics data for individuals who were targets of a marketing campaign. It has 42982 persons data and 367 features

```
In [191]: mailout_train.head()
```

Out[191]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	1763	2	1.0	8.0	NaN	NaN	NaN	NaN	8.0	15.0
1	1771	1	4.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
2	1776	1	1.0	9.0	NaN	NaN	NaN	NaN	7.0	0.0
3	1460	2	1.0	6.0	NaN	NaN	NaN	NaN	6.0	4.0
4	1783	2	1.0	9.0	NaN	NaN	NaN	NaN	9.0	53.0

5 rows × 367 columns

MAILOUT_TEST: Demographics data for individuals who were targets of a marketing campaign. It has 42833 persons data and 366 features

```
In [215]: mailout_test.head()
```

Out[215]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	1754	2	1.0	7.0	NaN	NaN	NaN	NaN	6.0	2.0
1	1770	-1	1.0	0.0	NaN	NaN	NaN	NaN	0.0	20.0
2	1465	2	9.0	16.0	NaN	NaN	NaN	NaN	11.0	2.0
3	1470	-1	7.0	0.0	NaN	NaN	NaN	NaN	0.0	1.0
4	1478	1	1.0	21.0	NaN	NaN	NaN	NaN	13.0	1.0

5 rows × 366 columns

Data Visualization

The below histograms show how the given data is distributed for the few features

Fig. 1 Following histogram shows how the best-ager typology is distributed

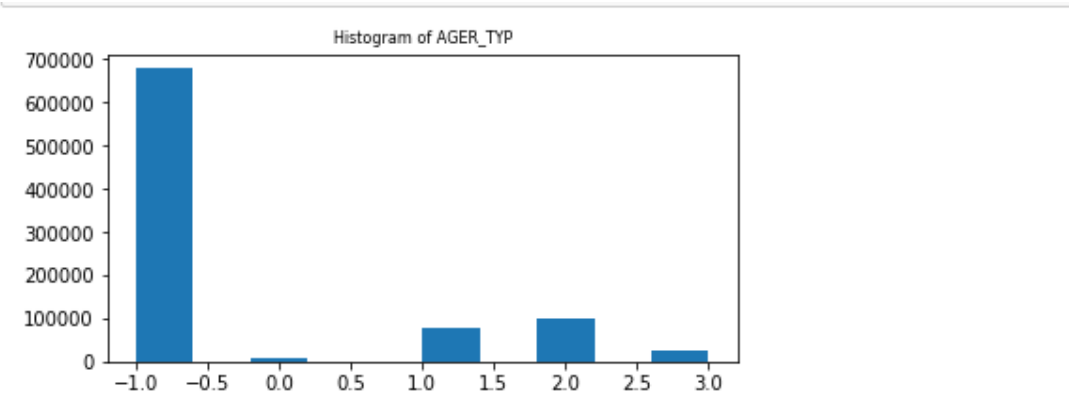


Fig. 2 Following histogram shows how the age classification through prename analysis is distributed

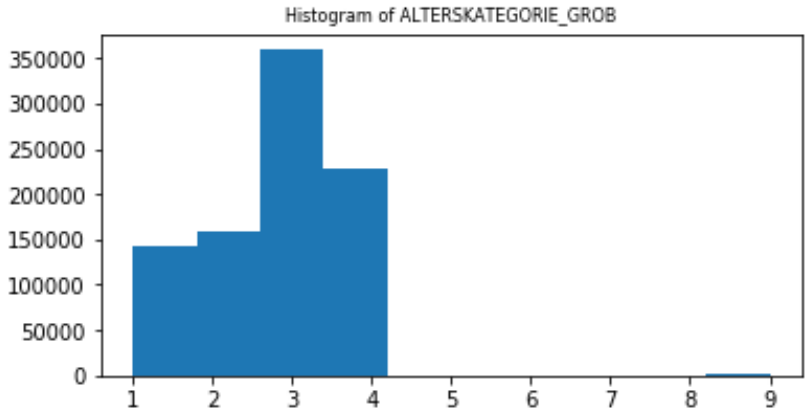


Fig. 3 Following histogram shows how the gender data is distributed

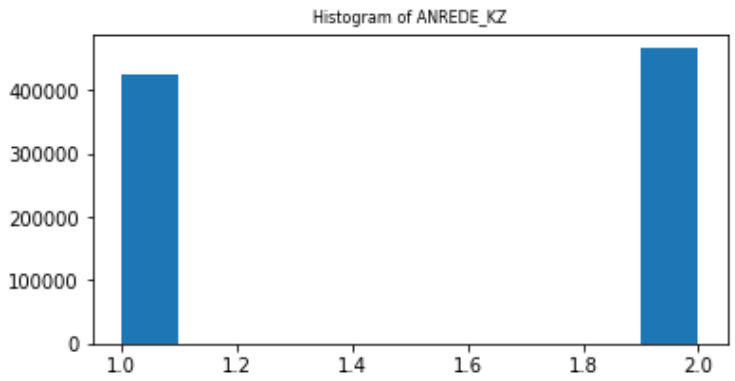


Fig. 4 Following histogram shows how the transaction activity BANKS in the last 24 months data is distributed

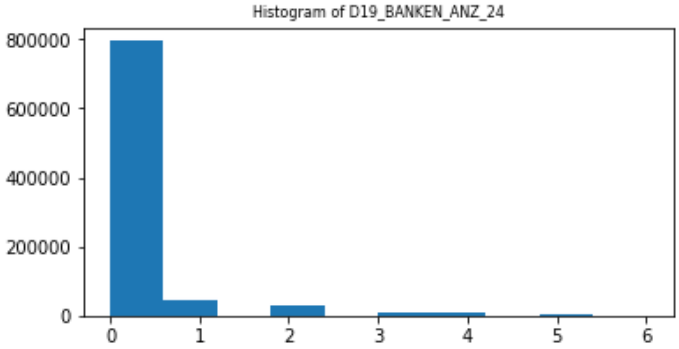
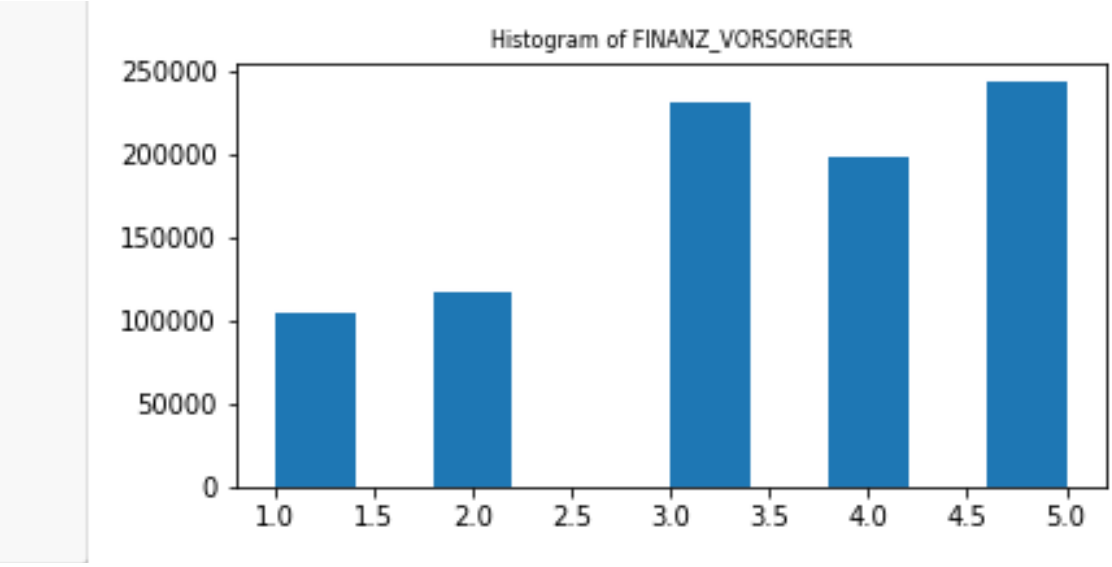


Fig. 5 Following histogram shows how financial typology be prepared is distributed



Benchmark

To create and initial benchmark for this imbalanced classifier, I am going to compare the Area under the receiver operating characteristic curve with a simple logistic regression score.

Methodology

Data Preprocessing

The preprocessing step consist of the following steps:

1. Load general population AZDIAS and customers data
2. Load and Prepare missing or unknown values based on given DIAS Attributes – Values 2017.xlsx.
3. Identify rows and columns which has missing or unknown data
4. Asses columns types
5. Prepare a clean function which removes and adds dummy columns

6. Impute the values for missing data
7. Apply feature scaling
8. Apply feature engineering e.g. Indicator variables (wealthy customer, age limit, type of cars, creating dummy variables etc.)

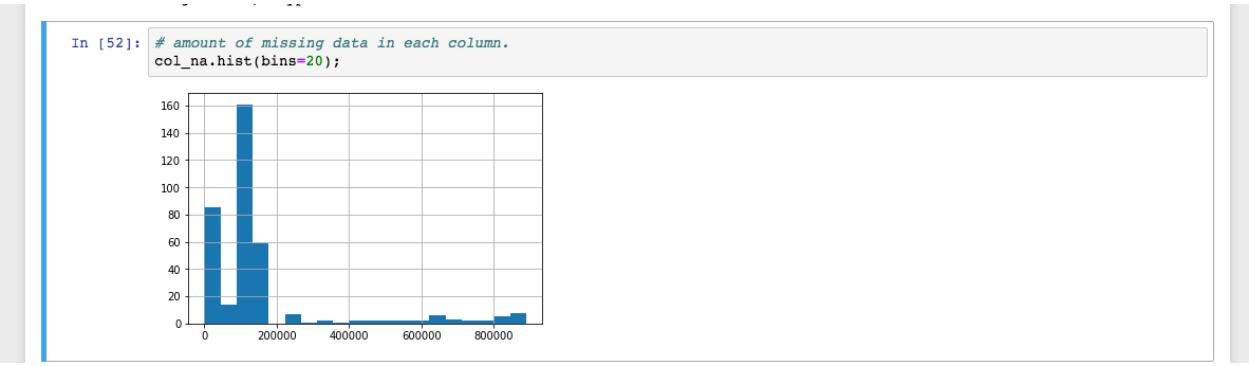
Implementation

I created a data frame of missing or unknown values based on DIAS Attributes – Values 2017.xlsx.

Prepared a missing or unknown (-1 or 0) columns list from the DIAS attributes sheet then compared with AZDIAS dataset features. Next find out the values that correspond to missing value codes of AZDIAS dataset and converted all of the missing or unknowns to NaNs.

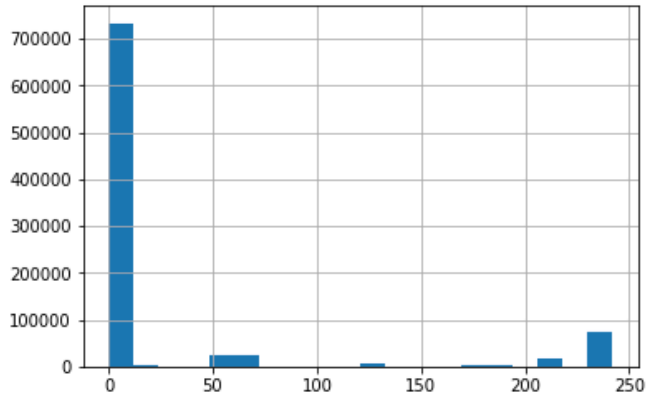
After analyzing the data, found that there are lot of rows and columns/features has missing data. The analysis identified that the columns with NaNs which were more than 200k and rows with NaNs which were more than 10 per row. Since these doesn't add any value for the analysis, these attributes were dropped from analysis.

Number of columns with NaNs



Number of Rows with NaNs

```
In [7]: # Plot histogram of number of missing values
row_na.hist(bins=20);
```



Additionally, there are other columns that were dropped based on the following reasons:

1. Column with unique values e.g LNR
2. Columns with more than 10 categories
3. Missing data e.g MIN_GEBAEUDEJAHR

By going through the DIAS attributes list, identified columns which are categorical e.g gender type

Crated dummy variables for columns with less than 10 unique values for simplicity. Adjusted feature OST_WEST_KZ data appropriately

At the end 732489 rows and 360 columns left.

After feature engineering, imputed missing values using mean method so that data doesn't have any more NANs.

Finally applied feature scaling so that principal component vectors are not influenced by the natural differences in scale for features.

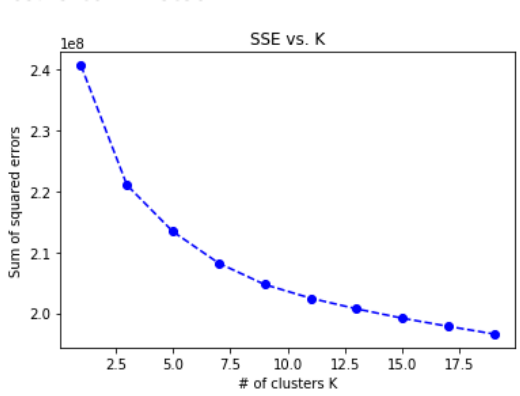
By this time, I have good data and ready to be fed into machine learning model. Each data point has 360 features, which means the data is 360 dimensional and an algorithm like K means has difficulties and result is often noisier clusters.

To address this issue, applied dimensionality reduction technique called Principal Component Analysis (PCA) on the dataset.

K-Means Clustering

Elbow graph method was used to identify an ideal number of clusters for KMeans clustering on the PCA-transformed data. Average of sum of squared errors (SSE) within-cluster distances was plotted against number of clusters.

The Elbow graph below indicates that good K is at 7.



Then unsupervised learning technique K Means clustering fit and predict methods applied to the general population and customers data. Then compare the proportion of data in each cluster for the customer data to the proportion of data in each cluster for the general population.

Supervised Learning Model

By this time, it is clear that which parts of the population are more likely to be customers of the mail-order company so build a prediction model. Load train data and implement pre-processing steps similar to customers dataset. Use various classifiers to train the dataset and select the model which has the highest ROC AUC score.

Kaggle Competition

Since model is already created and use it to predict which individuals are most likely to respond to a mailout campaign.

Refinement

To get the best result, run through the train data with various classifiers to select the highest ROC AUC classifier.

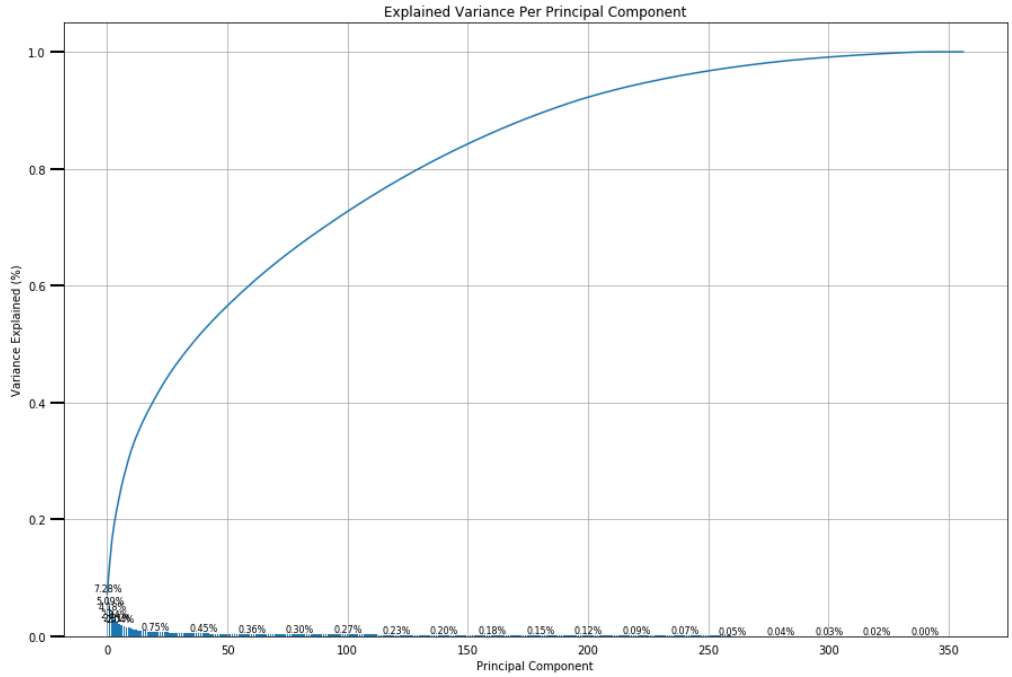
For the pre-processing step, we can use MICE for imputing unknown values and for the cluster quality use Shilhoutte scores.

Results

Model Evaluation and Validation

I plotted explained variance bar plot (see below). There is a visual reduction in explained variance after ~200 components. This number of transformed features results in 90% explained variance. So, 200 transformed features were retained for the clustering part of the project.

```
In [11]: scree_plot(pca_fit)
```



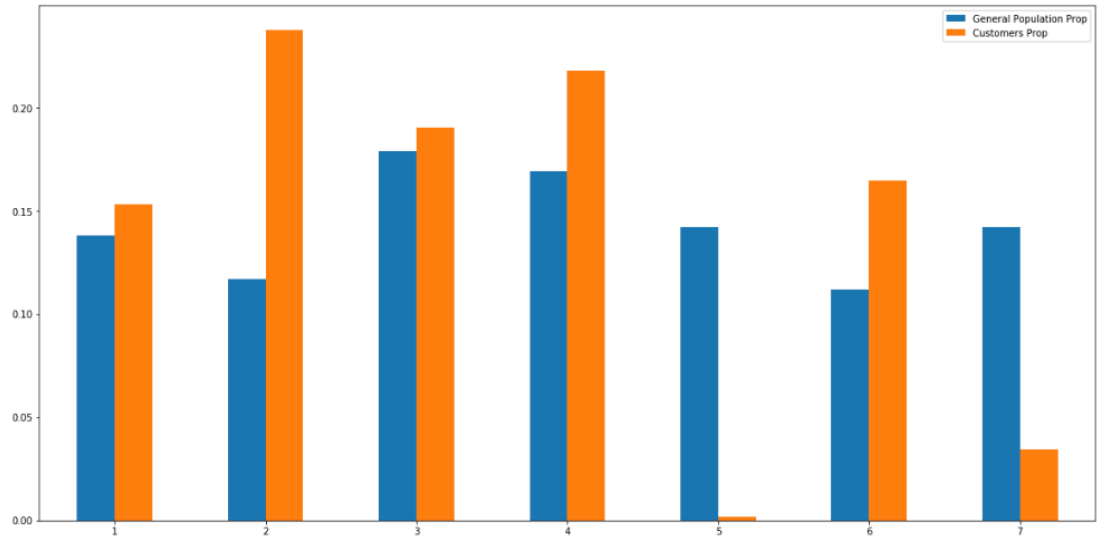
I chose to retain 200 components since it provides more than 90% of variance.

The highest ROC AOC is chosen because they performed the best among the tried combinations.

The results of clustering general population and customers data were compared to each other as shown in below Fig. Clusters with the highest positive difference in proportion between customers and general audience is overrepresented in the customers data (interested clusters #2 and #4)

```
In [76]: # Compare the proportion of data in each cluster for the customer data to the
         # proportion of data in each cluster for the general population.

         prop = pd.DataFrame({'General Population Prop': prop_cluster_general,
                              'Customers Prop': prop_cluster_customers})
         ax = prop.plot.bar(rot=0, figsize=(20, 10))
```



Clusters with the highest negative difference in proportion between customers and general audience are underrepresented in the customer data (no interest clusters #5 and #7).

Justification

The results of customer segmentation and ROC AUC scores proved that this model can be used for real world scenarios for any industry domain.

Conclusion

Reflection

The process used for this project can be summarized using the following steps:

- In the first part the assessment and preprocessing of the data was performed. There were 366 columns to analyze and not all of them had description. There were identified a lot of missing values and missing information about attributes. Feature engineering, Feature selection and handled missing data process was created that was further utilized in supervised and unsupervised parts.
- In the unsupervised part, the dimensionality reduction using PCA was performed to 200 latent features that describe 90% of explained variance. KMeans clustering to 7 clusters identified 2 clusters that are target customers of the company. These are share of midclass cars in the PLZ8 and 20–25 years of age.
- Lastly, choose Xgboost or Gradient Boosting Classifier and parameterize to build supervised model and make predictions over testing dataset on KAGGLE. The resulted performance of supervised learning algorithm is 70%.

This project potentially has some improvements. For example, there are other ways to preprocess the data: choose another threshold for dropping rows and columns, MICE, choose different transformations for the columns, apply MinMax Scaler instead of Standard Scaler, impute data in another way.

Improvement

Improvement of supervised model can be tested by using PCA dimensionality reduction. We could also choose attributes that have the most difference in clustering for overrepresented and underrepresented data and build supervised model using only these attributes. And also increase the hyper parameters for better optimization.